Scientometrics (2010) 85:257–270
DOI 10.1007/s11192-010-0160-5

Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature

Lawrence D. Fu · Constantin F. Aliferis

Received: 20 November 2009 / Published online: 3 February 2010

Abstract The most popular method for judging the impact of biomedical articles is citation count which is the number of citations received. The most significant limitation of citation count is that it cannot evaluate articles at the time of publication since citations accumulate over time. This work presents computer models that accurately predict citation counts of biomedical publications within a deep horizon of 10 years using only predictive information available at publication time. Our experiments show that it is indeed feasible to accurately predict future citation counts with a mixture of content-based and bibliometric features using machine learning methods. The models pave the way for practical prediction of the long-term impact of publication, and their statistical analysis provides greater insight into citation behavior.

Keywords Bibliometrics · Citation analysis · Machine learning · Information retrieval

Introduction

A commonly accepted method for measuring the impact of a scientific article is the *citation count*. This metric was first introduced by Gross and Gross (1927) and evaluates an article based on the number of citations it receives after publication. It is assumed that articles with higher impact receive more citations. The reasons for citing an article have been studied in detail, and today it is well understood that a citation does not necessarily endorse the quality of a cited article. For example, a citing article may refute a claim in a cited article, or the cited article may be non-essential to the citing article's hypothesis (Garfield 1962). Heuristic as citation count may be, it is still a useful measure of significance for the published literature. Despite its widespread use, it has some limitations. For instance, citation count should not be used to compare articles from different fields. Citations

L. D. Fu (✉) · C. F. Aliferis
Center for Health Informatics and Bioinformatics, New York University Medical Center,
333 E. 38th St, 6th Floor, New York, NY 10016, USA
e-mail: lawrence.fu@nyumc.org

accumulate at varying rates in different fields, and it is easier to receive citations in fields where articles are published more frequently or contain more references (MacRoberts and MacRoberts 1996; Seglen 1998; Phelan 1999).

We performed a broad survey of the citation analysis literature to estimate its size and contents. Google Scholar yielded approximately 23,100 articles for the query "citation analysis", and MEDLINE returned 3,108 articles for the query '"bibliometrics" [MeSH Terms] OR "bibliometrics" [All Fields]'. MEDLINE is a bibliographic database of articles published in the biomedical literature, and it is maintained by the National Library of Medicine. The vast majority of these articles focus on citation analysis and citation behavior instead of citation count prediction. We found only 12 articles that discuss citation prediction in the sciences. Notably the consensus in this literature is that predicting citations is extremely difficult if not impossible due to various discussed reasons (Getoor 2003; Rattigan and Jensen 2003; Feitelson and Yovel 2004). Lokker et al. (2008) recently presented a moderately predictive regression model for predicting citation counts 2 years after publication. This model used information available within 3 weeks after publication, and its model inputs were 17 article specific features and 3 journal specific features. Nine article-specific predictors were found to be statistically significant: the number of authors, clinical relevance score, number of pages, and number of references as well as if it was abstracted in an evidence-based medicine journal, contained a structured abstract, was an original article, was a multi-centered study, and was a study about therapy. Statistically significant journal specific predictive features were the number of databases that indexed the journal and the proportion of articles that were abstracted. The model captured 56% of the variation in a holdout test set ($R^2 = 0.56$). Sensitivity and specificity values were computed by identifying the top half or top third of cited papers. For the top half, the sensitivity and specificity of the model were 83.3 and 71.5%. For the top third, these values were 66.1 and 82.2%. The area under the receiver operating characteristic curve (AUC) was 0.76 for a median threshold of seven citations.

In a recently published report (Fu and Aliferis 2008), we presented early results demonstrating the feasibility of citation prediction models that satisfy previously unattainable properties. Specifically these models were highly predictive and used only information available at the time of publication. Most importantly, they were able to predict citation counts in a "deep" time horizon of 10 years. Our approach takes advantage of very recent developments in high-dimensional statistical pattern recognition techniques that can simultaneously analyze the full-text of articles and abstracts, yield high predictivity, and automatically identify important predictor variables out of thousands of candidate features (Burges 1998). We also incorporated into the model many predictive features that were unexamined in prior literature.

The present manuscript extends our preliminary experiments (Fu and Aliferis 2008) in four important ways. First, we perform a prospective independent validation. Model performance is evaluated on an independent set of articles that were published after the training set of articles. This analysis tests the robustness of the models for subsequent years and detects temporal variation in citation patterns. Second, we re-analyze our data with another modeling technique to determine whether the choice of classifier or features is responsible for improving predictivity over what has been achieved in the past literature. Third, we report the top predictive features for all thresholds and discuss the important issue of whether authors can manipulate such predictive models to increase citations to their articles (i.e., "game the models"). Fourth, we examine feature patterns to identify commonalities in highly-cited versus low-cited articles.

Methods

The prediction models are trained on two types of input features. Content features include the title, abstract, and MeSH terms. MeSH terms are medical subject heading terms used in MEDLINE to provide information about the topic of an article. Bibliometric features measure the quality of the authors, journals, and institutions. They include information such as the number of publications and citations for the first and last author in the previous 10 years as well as the journal impact factor. The full list of features is shown in Table 1.

We developed binary prediction models rather than continuous ones since error metrics for continuous loss functions (e.g., mean square error or percent of variation explained) are difficult to interpret in terms of their practical significance. The models predicted whether an article would exceed *T* citations within 10 years of publication where *T* is a threshold chosen prior to fitting models. The set of thresholds included 20, 50, 100, and 500 citations. These thresholds represent increasing levels of article impact from minor impact to strong impact, and these designations are relevant for a wide range of biomedical topics.

The corpus was designed by specifying a set of topics, journals, and dates. Eight topics were chosen randomly in internal medicine from the MeSH vocabulary tree: Cardiology, Endocrinology, Gastroenterology, Hematology, Medical Oncology, Nephrology, Pulmonary Disease, and Rheumatology. An article was operationally considered relevant to a topic if its MEDLINE record contained one of the eight MeSH terms, a related topic from the “See Also” field of the MeSH record, or a term from a sub-tree of one of these terms. Articles were included from a set of six journals with a broad range of impact factors: American Journal of Medicine, Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association, Lancet, and New England Journal of Medicine. Articles published between 1991 and 1994 were included so that citation data for a 10 year period after publication would be available. The window length provided sufficient time for citation rates to stabilize and thus reflect long-term impact. MEDLINE was queried for articles in the chosen journals and time frame, and information for additional predictive variables was downloaded from ISI Web of Science (WOS) at http://www.isiknowledge.com and the Academic Ranking of World Universities (ARWU) at http://www.arwu.org. The final corpus contained 3,788 documents, and the complete model consisted of 20,005 total features. Positive-to-negative class ratios for each

Table 1 List of input training features

Feature
Article title
Article abstract
MeSH terms
Number of articles for first author
Number of citations for first author
Number of articles for last author
Number of citations for last author
Publication type
Number of authors
Number of institutions
Journal impact factor
Quality of first author’s institution

threshold were as follows: 2705/1083 for threshold 20, 1858/1930 for threshold 50, 1136/2652 for threshold 100, and 100/3688 for threshold 500 citations.

Articles were formatted for learning by text preprocessing and term weighting. The title, abstract, and MeSH terms were extracted from MEDLINE records. PubMed stop words were removed from the title and abstract. Multiple forms of the same word were eliminated with the Porter stemming algorithm (Porter 1980) to reduce the dimensionality of the input space. Terms were weighted using log frequency with redundancy which considers term frequency in a document and the corpus (Leopold and Kindermann 2002; Aphinyanaphongs et al. 2005). Each weight was a value between 0 and 1. In the end, the corpus was represented as a matrix where rows corresponded to documents and columns represented terms. Bibliometric features were also scaled linearly between 0 and 1.

The learning algorithm was support vector machines (SVMs) with heterogeneous polynomial kernel (Leopold and Kindermann 2002; Aphinyanaphongs et al. 2005). This class of statistical machine learning methods has several desirable properties for predictive modeling: (a) they are well-regularized (i.e., penalize excessive variables) and thus resist overfitting; (b) the number of free parameters never exceeds the number of sample size which further resists overfitting; (c) SVMs are able to model very complex functions and perform efficient computational processing of non-linearities (i.e., higher-order interaction effects) by implicitly mapping to higher-dimensional spaces; (d) in empirical terms, SVMs have been very successful in many studies of text categorization for biomedical articles and web sites (Leopold and Kindermann 2002; Aphinyanaphongs et al. 2005).

Model selection was performed with a nested stratified n-fold cross validation design (Aliferis et al. 2006). SVM cost and polynomial kernel degree parameters were optimized in the inner loop (i.e., between training and validation parts of the data), and error estimation was performed in the outer loop with the independent data (i.e., in the remaining data which is the testing subset). The outer loop thus produces an unbiased estimate of model predictivity within each fold. The final estimate is averaged over all folds to reduce error estimate variance from the randomized data splits during training, validation, and testing. More details about using nested cross validation for analyzing high-dimensional data can be found in (Aliferis et al. 2006). Predictivity was measured using the area under the receiver operating characteristic curve (AUC).

After fitting the models and estimating their performance, we identified the most influential features with feature selection. We reduced the total number of features by selecting only features in the Markov Blanket of the response variable (i.e., number of citations received). The Markov Blanket is the set of features conditioned on which all remaining features are independent of the response variable. It excludes both irrelevant and redundant variables without compromising predictivity, and it provably yields maximum variable compression under broad distributional assumptions (Aliferis et al. 2009). The specific algorithm was semi-interleaved HITON-PC without symmetry correction which is an instance of the Generalized Local Learning class of algorithms (Aliferis et al. 2009). Logistic regression was performed on the selected features to estimate the magnitude of each feature's effect and statistical significance on predicting citation counts while controlling for other features in the logistic regression model.

We performed three additional analyses to extend our previous work. First, independent prospective validation was performed to further investigate the models' performance for future unseen cases. Corpus articles were set aside for independent validation purposes, and the remaining articles were used to derive predictive models using the original nested cross-validation procedure described previously. Models were trained on articles from 1991 and 1992 while articles from 1993 and 1994 were excluded from model building and used only for evaluation. The second extension was to use a different learning method besides SVM models. If performance is similar when using different modeling techniques, this result would indicate that performance depends more on the choice of features rather than the choice of classifier. Instead of training SVM models, we used a logistic regression based classifier along with the HITON-PC selected features. The third extension focused on the bibliometric features by learning decision trees. This step provided another method for identifying important patterns of features among the highly-cited articles.

Results

Main findings

Figure 1 shows that model predictivity ranged from AUC of 0.86 to 0.92. We note that an AUC of 0.80 indicates a highly predictive classifier. The SVM models accurately predicted whether a publication received a given number of citations for each citation threshold, and threshold 500 produced the most predictive model. This result demonstrates that it is possible to build models with high predictivity for long time horizons using only information at publication.

Prospective validation

Models were trained on articles from 1991 and 1992 while articles from 1993 and 1994 were excluded from model building and used for evaluation. Potential differences in predictivity between the two time periods is affected by two factors: (a) whether overfitting occurs during model selection and error estimation and/or (b) if there is temporal variability in citation patterns. We performed a response variable random reshuffling experiment borrowed from state-of-the-art analysis of high-throughput data (Aliferis et al. 2006) to investigate whether overfitting was present. Citation counts were randomly reshuffled, and models were rebuilt on the reshuffled data exactly as was done for non-shuffled data. This procedure yielded a distribution of AUC values that overlapped with 0.5 which established that the nested cross-validation model selection protocol is indeed unbiased in

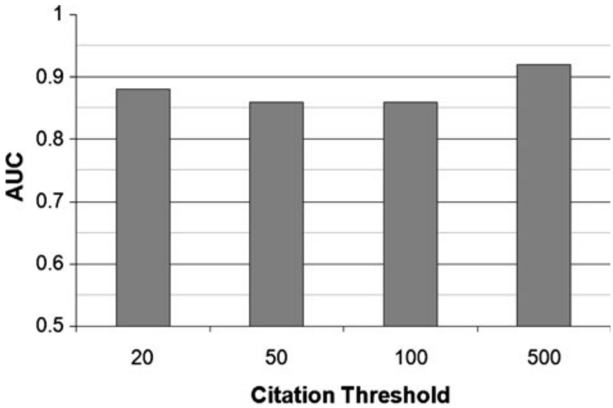

Fig. 1 AUC performance of citation count prediction models

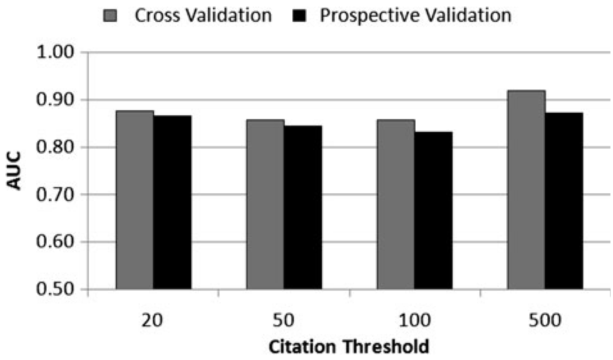

Fig. 2 Prospective validation results shown with cross validation results

our experiments (consistent with the theoretical properties of this protocol). Therefore the prospective validation informs us about the stability of citation patterns. Figure 2 shows the prospective validation results along with the original cross-validation results. For all thresholds except for 500, the cross-validation estimates were within the 98% confidence intervals for the prospective validation results. For threshold 500, the distribution of labels was very skewed with only 100 positive cases compared to 3688 negative cases. There would likely be less variation with more positive samples. The prospective validation results demonstrate that the prediction models can predict citation counts for future unseen articles.

Reduced models and feature importance

Feature selection was performed to simplify models and find the most predictive features. The total number of features was reduced to the Markov Blanket of the response variable (i.e., number of citations received) by applying the HITON-PC algorithm (Aliferis et al. 2009), and logistic regression was performed on the selected features. The original set of 20,005 features was reduced to 169, 125, 132, and 138 features for thresholds 20, 50, 100, and 500, respectively. Tables 2, 3, 4 show the top 10 ranked features for citation thresholds 20, 50, and 100. The results for threshold 500 are excluded since they were unreliable due to insufficient sample size. There were only 100 positive cases for 138 features.

A positive unit change in regression coefficient β for a feature corresponds to an increase of e^{β} in the odds of exceeding the citation count threshold for which the model is built. For example, "First Author Citations" had the largest coefficient of 5.75 for citation threshold 100. This value indicates that an article with the greatest number of first author citations was about 315 times ($e^{5.75} \approx 315$) more likely to receive 100 citations than an article with no first author citations (notice that a one-unit change for interval-based features corresponds to a difference between the largest and smallest values since interval variables were scaled to the [0,1] range).

This feature-specific analysis points to several important conclusions: (a) certain topics were associated with high citation rates (i.e., "hot topics" e.g., a paper discussing smoking mortality was 68 times more likely to exceed 100 citations than an article that did not discuss this topic when controlling for the other predictive variables); (b) citation history of authors played a significant role in citation rates by increasing the chances of exceeding

Table 2 Top 10 features sorted by absolute value of regression coefficient (threshold 20)

Feature	Regression coefficient	*P* value	Standard error
Cardiac tamponade [MeSH]	−4.94	0.000	1.28
Splenomegali	−4.94	0.007	1.83
Journal impact factor [WOS]	4.04	0.000	0.25
Supply and distribution [MeSH]	−3.97	0.002	1.26
Ectopi	−3.59	0.007	1.32
Thrombocytopenia:immunology [MeSH]	−3.56	0.008	1.34
Internal medicine [MeSH]	−3.54	0.001	1.02
Lung neoplasms:etiology [MeSH]	−3.44	0.001	1.00
Cholelithiasis [MeSH]	−3.27	0.010	1.27
Kidney failure, chronic:metabolism [MeSH]	−3.11	0.004	1.09

Table 3 Top 10 features sorted by absolute value of regression coefficient (threshold 50)

Feature	Regression coefficient	*P* value	Standard error
Splenectomi	−3.41	0.006	1.24
Journal impact factor [WOS]	3.34	0.000	0.16
Last author citations [WOS]	3.15	0.001	0.91
Ciprofloxacin	−2.86	0.019	1.22
Anemia, sickle cell [MeSH]	−2.76	0.000	0.68
Rural health [MeSH]	−2.67	0.015	1.10
Brain	2.57	0.000	0.64
History [MeSH]	−2.44	0.046	1.23
Zidovudine:therapeutic use [MeSH]	2.42	0.030	1.11
Death, sudden [MeSH]	−2.33	0.014	0.95

Table 4 Top 10 features sorted by absolute value of regression coefficient (threshold 100)

Feature	Regression coefficient	*P* value	Standard error
First author citations [WOS]	5.75	0.000	1.47
Smoking:mortality [MeSH]	4.22	0.018	1.79
Offset	3.35	0.007	1.23
Journal impact factor [WOS]	3.32	0.000	0.18
Last author citations [WOS]	3.02	0.001	0.87
Birth weight [MeSH]	2.95	0.000	0.77
Pilot projects [MeSH]	−2.91	0.013	1.17
Autoantibodies:blood [MeSH]	2.78	0.001	0.81
Family practice [MeSH]	−2.75	0.016	1.14
gy	2.65	0.006	0.96

100 citations by 315 times when comparing the best and worst author citation histories; (d) For each threshold, different sets of content features were selected (and ranked differently in the top positions) which indicates that various topics strongly predict distinct levels of citation impact consistent with prior literature that demonstrated different citation patterns in different fields of medicine (MacRoberts and MacRoberts 1996; Seglen 1998; Phelan 1999). On the other hand, bibliometric features and impact factor are predictive independent of threshold and always have large positive effects for all thresholds studied.

A heatmap was created in Table 5 to visually display the relative importance of the features. Blank entries indicate that a feature was not selected with HITON for a given threshold and thus excluded from logistic regression. Journal impact factor was the only feature that ranked highly for all three thresholds. Other features that were important for multiple thresholds were "Last Author Citations", "Pilot Projects [MeSH]", and "History [MeSH]".

Manipulation of models

It is conceivable that authors aware of which features correlate with highly cited papers could try to take advantage of this information by "gaming" or manipulating the models. In other words, authors could attempt to artificially inflate the citation counts of their articles by tailoring their articles towards features that correlate with high citation counts. Examining the features shows that manipulating the models is very difficult in practice, if not impossible. For example, authors could insert content terms that occur in highly-cited papers into their articles. However, if the paper's original topic focus or contents were not related to these added terms, reviewers and editors would easily recognize this attempt during the review process. In terms of the bibliometric features, it is similarly impossible for an author to artificially inflate the features without improving the quality of the work. An author cannot publish an article in a high impact factor journal, convince primary authors with strong citation histories to collaborate, or change the citation history of primary authors for otherwise poor articles.

Additional modeling method

We used an alternate modeling technique besides SVM models to investigate whether performance was affected more by the choice of classifier or features. The original performance with the SVM classifier was compared to a logistic regression based classifier using the HITON-PC selected features. Table 6 shows that the original SVM classifier results were slightly higher than the logistic regression based classifier. The regression based classifier still performed as a relatively good classifier. The comparable performance with different classifiers shows that the choice of features is more important than the choice of classifier for prediction purposes.

Bibliometric feature patterns

Decision tree learning was performed on the bibliometric features to extract feature patterns. Bibliometric features were converted into binary values prior to decision tree learning to yield trees that are easier to interpret, and the trees are shown in Figs. 3, 4, 5. For a given feature, the left branch indicates an article with a value in the lower half of values for the corpus. The right branch represents the top half of values in the corpus. For example, the left branch for "Last Author Citation Count" corresponds to articles in the bottom half of the range of values for last author citation count in the corpus. For “Article Type”, the left branch corresponds to papers while the right branch represents review articles. A leaf value less than 0.5 represents predicting an article will receive less than the threshold number of citations. For example, the tree for threshold 100 in Fig. 5 shows that an article with a high impact factor, a high quality institution, and a last author with many citations would be predicted to receive at least 100 citations. This classification is shown by following the right

Table 5 Heatmap of logistic regression coefficients

Feature Name	Threshold		
	20	50	100
First Author Citations [WOS]			5.75
Cardiac Tamponade[MeSH]	4.94		
splenomegali	4.93		
Smoking:mortality[MeSH]			4.22
Journal Impact Factor [WOS]	4.04	3.34	3.32
supply & distribution[MeSH]	3.97		
ectopi	3.59		
Thrombocytopenia:immunology[MeSH]	3.56		
Internal Medicine[MeSH]	3.54		
Lung Neoplasms:etiology[MeSH]	3.44		
splenectomi		3.41	
offset			3.35
Cholelithiasis[MeSH]	3.27		
Last Author Citations [WOS]		3.15	3.02
Kidney Failure, Chronic:metabolism[MeSH]	3.11		
Ventricular Fibrillation[MeSH]	2.96		
Birth Weight[MeSH]			2.95
tomographi[Title]	2.94		
Pilot Projects[MeSH]		1.36	2.91
increment	2.89		
ciprofloxacin		2.86	
Autoantibodies:blood[MeSH]			2.78
gradual	2.77		
Anemia, Sickle Cell[MeSH]		2.76	
Family Practice[MeSH]			2.75
history[MeSH]	2.69	2.44	
Rural Health[MeSH]		2.67	
Oxygen:blood[MeSH]	2.66		
gy			2.65
tachycardia[Title]	2.58		
person[Title]			2.58
brain		2.57	
periton[Title]	2.48		
Mycobacterium tuberculosis[MeSH]			2.47
tran			2.46
Zidovudine:therapeutic use[MeSH]		2.42	
clinicopatholog[Title]	2.42		
Immunohistochemistry[MeSH]			2.37
Death, Sudden[MeSH]		2.33	
Endothelium, Vascular[MeSH]			2.26
pylori		1.52	2.25

Table 5 continued

Feature Name	Threshold		
	20	50	100
catecholamin		2.21	
uncompl		2.17	
hypoglycaem		2.14	
Clinical Protocols [MeSH]	2.10		
sucralf	2.03		
meta[Title]			1.95
quantifi			1.88
inappropri		1.86	
Kidney Diseases[MeSH]			1.84
european[Title]	1.81		
transmiss	1.79		
present[Title]	1.79		
ambulatori[Title]		1.78	
took		1.71	
liver[Title]	1.64		
apolipoprotein			1.60
Molecular Sequence Data[MeSH]		1.59	0.82
shift	1.58		
Atrial Fibrillation [MeSH]		1.57	
mutat[Title]			1.54
heparin			1.53
output		1.52	
Decision Making[MeSH]	1.48		
Article Type [WOS]	1.15	1.48	
unselect			1.40
chain		1.30	
concomit	1.23		
endogen			1.22
largest			1.18
spontan			1.17
thrombosi		1.08	
endoscop			0.98
asthma		0.96	
gener[Title]		0.92	
conclus	0.91	0.73	
larger		0.89	
circul		0.87	
syndrom	0.84		
Acute Disease[MeSH]		0.76	0.72
trial[Title]			0.75
adjust	0.72		
neg	0.71		
cumul			0.71
abov		0.68	
defin	0.67		
variabl	0.67		
mediat			0.66

Table 5 continued

Feature Name	Threshold		
	20	50	100
adult	0.54		
specif			0.53
estim		0.53	
symptom	0.52		
ratio		0.41	0.52
myocardi			0.46
surgeri		0.45	
combin		0.45	
random			0.41
associ	0.40	0.33	
low	0.40	0.29	
occur	0.38		
Middle Aged[MeSH]		0.38	
particip			0.37
mortality[MeSH]			0.37
Quality of Institution [WOS]	0.33	0.36	0.31
p		0.32	0.35
trial		0.35	
diseas	0.35		
factor		0.31	
complications[MeSH]		0.29	
mortal		0.28	
month		0.27	
relat		0.24	
year		0.23	

Table 6 Performance of original SVM classifier compared to logistic regression based classifier

Threshold	SVM classifier AUC	Logistic regression based classifier AUC
20	0.88	0.80
50	0.86	0.78
100	0.86	0.77
500	0.92	0.82

branch at every feature starting from "Impact Factor" through "Quality of Institution" and "Last Author Citation Count" before reaching a leaf value of 1.

The tree for threshold 100 in Fig. 5 shows other interesting feature patterns. If an article was published in a high impact journal but originated at a lower quality institution, it could still receive more than 100 citations if the number of authors was high or the authors had high citation counts. If an article was published in a journal with a low impact factor, it was

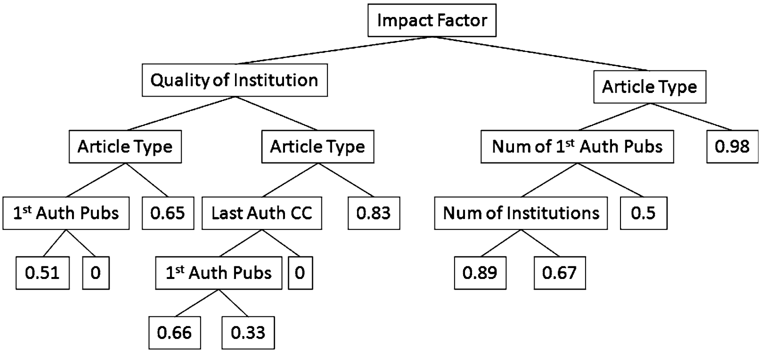

Fig. 3 Decision tree learned from bibliometric features (threshold 20)

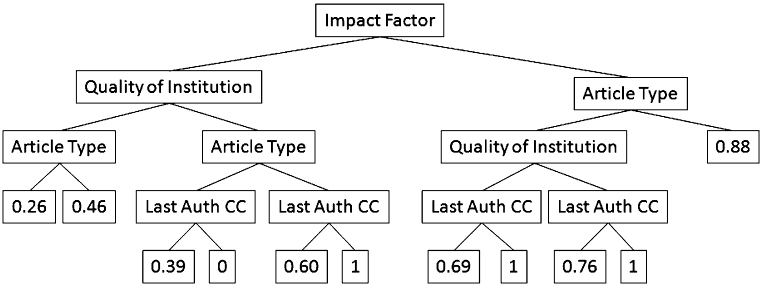

Fig. 4 Decision tree learned from bibliometric features (threshold 50)

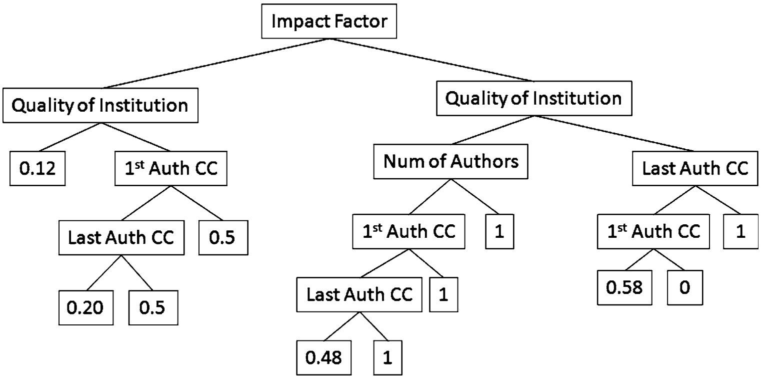

Fig. 5 Decision tree learned from bibliometric features (threshold 100)

less likely receive 100 citations since the largest leaf value was 0.5 regardless of the other features. The decision tree for threshold 50 in Fig. 4 similarly shows the importance of impact factor. The right subtree for articles with high impact factor contains only leaf values greater than 0.5.

Discussion

Our results show that it is feasible to predict future citation counts of an article immediately upon publication with a deep time horizon and with high predictivity. Also, the data sheds light on influential features that correlate strongly with and may determine future citation rates. Compared to Lokker's work (Lokker and Mckibbon 2008), we not only included information about the journal and article but also incorporated information about the citation history of authors and the quality of their institutions. Moreover, our features are automatically generated and do not require human raters to manually provide scores of clinical relevance and newsworthiness for each article as Lokker's method does. Our approach therefore requires less manual labor which makes it easier to extend and apply to other areas of biomedicine.

Our earlier work (Fu and Aliferis 2008) demonstrated the technical feasibility of the learning approach, and experiments in the present manuscript extend the work in four important ways. First, we rebuilt models and performed a prospective independent validation that shows the robustness of the models in subsequent years. In contrast, the authors of (Lokker and McKibbon 2008) did not perform prospective validation, and it is unknown how well their model will perform on future cases. Second, we re-analyzed our data using a logistic regression based classifier and found that the choice of classifier for this dataset did not significantly affect the performance of the obtained models. Thus we conclude that it is the choice of input features that results in high predictivity. We also publish here for the first time the top ranking features, explain why it would be difficult to manipulate the prediction models, and extract prominent feature patterns of highly and low-cited articles.

At least two important related questions are open research problems at this time. First, are the identified features causally affecting future citations or are they confounded by hidden citation determinants? Secondly, are the causal factors a result of some type of bias, or are they legitimate inducers of future citations? Future work will address these questions.

In conclusion, our data show that the development and validation of practical citation prediction tools for use by researchers and clinicians is a realistic possibility. Such tools have the potential to render citation counts a more practical tool for evaluating long-term impact of recent work and their authors. A lot remains to be learned about what causes articles to be highly cited and how to introduce such models in everyday literature search and evaluation.

References

Aliferis, C., Statnikov, A., et al. (2009). Local causal and markov blanket induction for causal discovery and feature selection for classification. *JMLR* (accepted).

Aliferis, C., Statnikov, A., et al. (2006). Challenges in the analysis of mass-throughput data. *Cancer Informatics, 2*, 133–162.

Aphinyanaphongs, Y., Tsamardinos, I., et al. (2005). Text categorization models for high-quality article retrieval in internal medicine. *JAMIA, 12*(2), 207–216.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Feitelson, D., & Yovel, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of Documentation, 60*(1), 44–61.

Fu, L., & Aliferis, C. (2008). *Models for predicting and explaining citation count of biomedical articles*. AMIA symposium.

Garfield, E. (1962). Can citation indexing be automated? *Essays of an Information Scientist, 1*, 84–90.

Getoor, L. (2003). Link mining: A new data mining challenge. *SIGKDD Explorations, 5*(1), 84–89.

Gross, P., & Gross, E. (1927). College libraries and chemical education. *Science, 66*, 385–389.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. *Machine Learning, 46*, 423–444.

Lokker, C., McKibbon, K. A., et al. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *BMJ*. http://www.bmj.com/cgi/content/abstract/bmj.39482.526713.BEv526711.

MacRoberts, M., & MacRoberts, B. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435–444.

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117–136.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*, 130–137.

Rattigan, M., & Jensen, D. (2003). The case for anomalous link discovery. *SIGKDD Explorations, 5*(1), 41–47.

Seglen, P. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica, 69*(3), 224–229.